Theor Appl Genet (2014) 127:981–994
DOI 10.1007/s00122-014-2273-3

ORIGINAL PAPER

# Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population

**David Cros · Leopoldo Sánchez · Benoit Cochard · Patrick Samper · Marie Denis · Jean-Marc Bouvet · Jesús Fernández**

Received: 21 March 2013 / Accepted: 22 January 2014 / Published online: 7 February 2014


**Abstract**
***Key message*** **Explicit pedigree reconstruction by simulated annealing gave reliable estimates of genealogical coancestry in plant species, especially when selfing rate was lower than 0.6, using a realistic number of markers.** Genealogical coancestry information is crucial in plant breeding to estimate genetic parameters and breeding values. The approach of Fernández and Toro (Mol Ecol 15:1657–1667, 2006) to estimate genealogical coancestries from molecular data through pedigree reconstruction was limited to species with separate sexes. In this study it was extended to plants, allowing hermaphroditism and monoecy, with possible selfing. Moreover, some improvements were made to take previous knowledge on the population demographic history into account. The new method was validated using simulated and real datasets. Simulations showed that accuracy of estimates was high with 30 microsatellites, with the best results obtained for selfing rates below 0.6. In these conditions, the root mean square error (RMSE) between the true and estimated genealogical coancestry was small (<0.07), although the number of ancestors was overestimated and the selfing rate could be biased. Simulations also showed that linkage disequilibrium between markers and departure from the Hardy–Weinberg equilibrium in the founder population did not affect the efficiency of the method. Real oil palm data confirmed the simulation results, with a high correlation between the true and estimated genealogical coancestry (>0.9) and a low RMSE (<0.08) using 38 markers. The method was applied to the Deli oil palm population for which pedigree data were scarce. The estimated genealogical coancestries were highly correlated (>0.9) with the molecular coancestries using 100 markers. Reconstructed pedigrees were used to estimate effective population sizes. In conclusion, this method gave reliable genealogical coancestry estimates. The strategy was implemented in the software MOLCOANC 3.0.

Communicated by M. Frisch.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-014-2273-3) contains supplementary material, which is available to authorized users.

D. Cros (✉) · B. Cochard · P. Samper · M. Denis · J.-M. Bouvet
Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit (AGAP), CIRAD, International campus of Baillarguet, TA A-108/C, 34398 Montpellier Cedex 5, France
e-mail: david.cros@cirad.fr

L. Sánchez
Forest Tree Improvement, Genetics and Physiology Research Unit (AGPF), INRA, 2163 Avenue de la Pomme de Pin, CS 40001 Ardon, 45075 Orleans Cedex 2, France

J. Fernández
Departamento de Mejora Genética Animal, INIA, Ctra. Coruña Km 7.5, 28040 Madrid, Spain

## Introduction

Knowledge of genetic relationships between individuals in plant populations is of major interest for breeding as it allows estimation of the genetic parameters (heritabilities and genetic correlations), breeding values and effective population sizes. Moreover, this information is needed to reduce the increase of inbreeding and loss of diversity through the control of global coancestry. These values are conventionally obtained from genealogical information.

However, pedigrees of breeding populations of plant species may be completely, or at least partially, unknown

for various reasons. Breeding of many plant species includes uncontrolled open pollination or controlled pollination with a pollen mixture, for instance to carry out open pollinated or polycross progeny tests, as in forage crops and forest trees. Missing genealogical information could also be due to the existence of cutoff dates for recording the pedigree (i.e., pedigree recording does not start from the unrelated founder individuals) and long breeding cycles. Furthermore, breeding programs include many error-prone steps, including pollination, seed preparation, germination, planting, etc. Mistakes may occur even for controlled crosses between clearly identified parents, leading to illegitimate progenies (one or both parents actually unknown) or contamination (some illegitimate individuals in the progeny). This has been reported many times, for instance in sugarcane (McIntyre and Jackson 2001), oil palm (Corley 2005), Douglas-fir and loblolly pine (Adams et al. 1988). In addition, individuals of unknown origin can enter the breeding population, such as when coming from another breeding program.

Completely missing genealogical information, small depth pedigrees or erroneous ones have negative effects on breeding programs, often due to erroneous estimations of genetic parameters and breeding values. In diallel progeny trials, Ericsson (1999) showed that 0.5 % of trees with misidentified pedigree were enough to downward bias additive variance and heritability estimates. Atkin et al. (2009) found that the accuracy of additive variance components and breeding values improved when more pedigree information was used in analyses. In polycross progeny tests, the unknown relative contribution of male parents can be detrimental (Kumar et al. 2007) and reconstructing male parentage increased the genetic gain (Doerksen and Herbinger 2010). Furthermore, assuming that individuals newly imported into a breeding population are unrelated to the main population may lead to erroneous management decisions if a common ancestral origin actually exists.

Therefore, there is substantial interest in having good knowledge of the pedigree of breeding populations of plant species. In particular, this would result in reliable estimates of the genealogical coancestry and, consequently, in better breeding management. Fernández and Toro (2006) proposed an approach (FT method, thereafter) for estimating the genealogical coancestry between contemporaneous genotyped individuals through the construction of an explicit virtual genealogy. Through a *simulated annealing* algorithm (Kirkpatrick et al. 1983), FT finds the pedigree that yields a genealogical coancestry matrix with the highest correlation relative to the actual molecular coancestry matrix, or any other provided relatedness matrix. In other words, FT matches molecular coancestries within a given current population to the most likely compatible pedigree. Several parameters help to reduce the space of parametric solutions and to find a solution closer to the true pedigree, like the number of previous discrete generations to reconstruct, the maximum number of sires and dams in those previous generations and, if available, a known part of the pedigree. The advantages of FT over other methods (see Blouin 2003; Butler et al. 2004; Pemberton 2008 for reviews) are that it does not require knowledge of the true allelic frequencies in the base population (i.e., founder individuals from which the genotyped individuals derived) and that Hardy–Weinberg and linkage equilibria in the base population are not necessary. Moreover, it can manage any degree of complexity in relatedness between individuals and always provides congruent relationships, resulting in positive definite pedigree-based coancestry matrices.

The goals of this study were first to extend the capabilities of the FT method, in order to make it suitable for plant species, and second to demonstrate these new capabilities with simulated and true plant breeding populations.

FT was developed for dioecious species, i.e., species with separate sexes, especially animals. Consequently, any virtual ancestor acting as a node in the simulated pedigree in FT was either a male or a female, never both simultaneously. Therefore selfing, which is possible in many plant species and a tool for their breeding, was not possible. We extended the FT approach to encompass monoecy and hermaphroditism, with the possibility of selfing (i.e., mixed-mating, where a fraction of the progeny is derived from self-fertilization and the remainder from outcrossing), either when mixed-mating is the natural mode of reproduction or because this is artificially forced for breeding purposes. The first generation with selfings is a user-defined parameter, thus accounting for the possibility of natural cross-fertilizing species that, at a particular time, entered a breeding program in which self-fertilization could be artificially conducted. Another limitation of the FT approach was that the size of the virtual population was constant over generations. In order to take knowledge of the population demographic history into account, we made FT able to consider a variable number of ancestors through generations. The possibility of starting from related founders was also implemented, as it helps to get solutions which fit better to the real data.

To demonstrate the capabilities of the new method (FT*, thereafter), we used simulated data of a mixed-mating species and real data from two oil palm (*Elaeis guineensis*) breeding populations. Regarding the simulated data, we evaluated the effects of different selfing rates, percentages of unknown parentages and numbers of markers on the accuracy of the method, as well as the effects of departures from the ideal situation of Hardy–Weinberg and linkage equilibria. The real data involved the Yangambi (Africa) and the Deli (Asia) oil palm populations, for which important breeding efforts are currently under way. The Yangambi

population was used to validate FT*, as the pedigree is well known back to founder individuals. For the Deli population, the pedigree data are scarce and we used this population to illustrate one key application of pedigree reconstruction with FT*, the estimation of the pedigree-based effective population size (*Ne*) via the approach developed by Gutiérrez et al. (2008, 2009) and Cervantes et al. (2011). *Ne* is a parameter of interest in oil palm and no estimates are available. This species is "temporally dioecious", producing male and female inflorescences in an alternating cycle on the same plant, resulting in an allogamous mode of reproduction. Consequently, selfing does not occur in nature, but it has been used by breeders at some point in its pedigree. The inference of the extent of man-made selfing events is therefore of importance for current breeding population management. The *Ne* values obtained with the reconstructed pedigree were compared to those obtained via the method of Hill (1981) and Waples (2006) which is based only on linkage disequilibrium and independent of pedigree data.

## Materials and methods

### Original algorithm and new additions

The original method of Fernández and Toro (2006) starts with the generation of a random pedigree for the genotyped individuals. Alternative solutions are generated by randomly substituting one of the ancestors for another of the same generation. Alternative solutions are checked to avoid incompatible full-sib families at the molecular level. Valid solutions are used to calculate the genealogical coancestry matrix of genotyped individuals with the tabular method (Emik and Terrill 1949) and its correlation with the molecular coancestry matrix, which is calculated according to Eding and Meuwissen (2001). Coancestry coefficients are defined as the probability that two alleles taken at random, one from each individual, are identical by descent (genealogical coancestry) or by state (molecular coancestry). The probability of acceptance of alternative solutions is a function of the difference in the correlation of genealogical coancestry with molecular coancestry between the alternative and current solution and the cooling factor. The optimal solution is reached when no alternative solutions are accepted for 5,000 changes at a given 'temperature' or when the maximum number of steps is performed.

The FT method was modified to include the following new features:

1. Monoecy and hermaphroditism were implemented in addition to dioecy, so that the algorithm can create virtual ancestors either with separate sexes or with both sexes at once.
2. The possibility of selfing was implemented, as an option within monoecy and hermaphroditism, from the base population (i.e., in the whole pedigree) or from a later user-specified generation. This latter possibility is relevant when known artificial self-fertilization has been recently started in a species with non-natural selfing. When selfing is allowed, the self coancestries are also included in the calculation of the correlation between molecular and estimated genealogical coancestry matrices. Furthermore, a new rule was added for selfing when checking Mendelian inheritance in initial and alternative solutions considered by the simulated annealing algorithm, as no more than two alleles could exist in a full-sib family arising from selfing. Modifications 1. and 2. allow for the application of the method to the different modes of sexual reproduction existing in plant species.
3. The maximum number of ancestors per generation can be separately defined by the user. In this way, any previous information on the demographic history of the population can be taken into account (e.g., number of founders, known bottlenecks or expansions). These explicit limits in the size of generations reduce the space of feasible solutions for the optimization process. Thus the optimal solution is found more easily by the algorithm and is expected to be closer to the true pedigree.
4. We also included the possibility of accounting for a predefined coancestry matrix between founders, whenever real knowledge on the origin of the oldest known ancestors is available or simply to compare different hypotheses about the foundation of the population (i.e., from related or unrelated individuals).

All the other features of the FT method (coping with genotyping errors, including known relationships, etc.) were included in the new version too. The original as well as the new code were written in FORTRAN and a compiled version of the software for Windows platforms can be freely downloaded (MOLCOANC version 3.0, at http://dl.dropbox.com/u/5714008/Fernandez.htm). In addition, pedigrees are now automatically drawn with 'Pedigraph' (Garbe and Da 2008) if this software is already installed in the computer.

### Testing the effect of selfing rate and marker numbers with simulated data

We simulated pedigrees of an expanding population of a hermaphroditic species, with different selfing rates and variable numbers of simple sequence repeat markers (SSR), in order to test the effects of these two parameters on the accuracy of genealogical coancestries estimated by FT*. The population started with five founders and reached 40 individuals in the sixth generation. In the founder

generation, 10, 30 or 90 SSR were simulated by randomly drawing without replacement alleles from a pool of 10 equiprobable alleles, independently for each marker. Consequently, the markers were expected to be in Hardy–Weinberg and linkage equilibria at the initial generation (base population). Notice that the sampling process produced data with a strong variation in allelic frequencies within loci and actual number of alleles between loci. This way simulated data better mimics the kind of scenarios which could be found in real data. Sequence repeat markers were evenly distributed along a genome of 10 chromosomes of 160 cM each, resulting in a genome length close to the 1.7 M found in oil palm (Billotte et al. 2005), in order to facilitate the comparison of results between simulations and real datasets. Individuals were considered as male and/or female and mated randomly, at first with the exclusion of selfing, while making sure that there was at least one mating per individual. Afterwards, six selfing rates were tested (0, 0.2, 0.4, 0.6, 0.8 and 1). The selfing rate was defined as the ratio of the number of individuals being the offspring of a self-fertilization to the total number of individuals (excluding founders). To achieve the defined selfing rates, a corresponding number of random crosses were converted into selfings. When the selfing rate differed from zero, selfing was allowed from the first generation and the same selfing rate was applied to each generation. Fifty pedigrees were simulated for each combination of selfing rate and number of SSR. Segregation of founder alleles in the pedigree was simulated with the R 'pedantics' package (Morrissey and Wilson 2010), with the mutation rate set at zero. FT* was applied to individuals of the last generation to reconstruct their pedigree from their molecular data and to estimate their genealogical coancestry.

For this and the following simulations, as well as for the real datasets, details about datasets and parameters used for the pedigree reconstruction algorithm are given in Table 1. For the *simulated annealing*, the maximum number of steps allowed was 150, the number of solutions tested at each step was 5,000, the initial temperature was 0.9 and the rate of temperature decrease was 0.99, in all situations.

**Table 1** Marker data, true pedigree data and control parameters for simulated annealing for each case studied

| Case study | | | | | |
|---|---|---|---|---|---|
| | Simulation | | | Oil palm real data | |
| | Selfing rate | % of unknown parentages | HW/LD | Yangambi | Deli |
| True pedigree data | | | | | |
| Number of generations | 6 | 5 | 6 | 5 | Unknown |
| Number of individuals per past generation[a] | 5, 10, 15, 20, 30 | 5, 10, 15, 20 | 20, 25, 30, 40, 50 | 9, 7, 7, 5 | Unknown |
| First generations with selfings allowed | 1 | 1 | 1 | 2 | Unknown |
| Marker data | | | | | |
| Number of SSR | 10, 30, 90 | 10, 36, 63, 90 | $92 \pm 6$, $103 \pm 8$[b,e] | 6–166 | 8–160 |
| Number of genotyped individuals[c] | 40 | 25 | 60 | 16 | 104 |
| Average number of alleles per SSR | $5.5 \pm 1.0$[b] | $5.4 \pm 1.0$[b] | $2.2 \pm 0.4$, $2.1 \pm 0.4$[b,e] | 3.6 (2–6)[d] | 2.7 (2–5)[4] |
| Control parameters for MOLCOANC | | | | | |
| Number of generations to reconstruct | 5 | 4 | 5 | 4 | 7–9 |
| Maximum number of individuals per reconstructed generation[c] | 10, 20, 30, 40, 60 | 10, 20, 30, 40 | 40, 50, 60, 80, 100 | 64, 64, 64, 32 | 4, 30, 60, 31, 19, 3, 8–4, 30, 30, 30, 60, 31, 19, 3, 8 |
| First generations with selfing allowed | 1 | 1 | 1 | 2 | 4–5 |

*HW* Hardy–Weinberg disequilibrium, *LD* linkage disequilibrium

[a] Starting from the founder generation

[b] Computed over all replicates (±SD)

[c] i.e., Individuals of the last generation

[d] Range for the studied dataset

[e] Low and disequilibriums conditions, respectively

### Testing the effect of percentage of unknown parentages and marker numbers with simulated data

These simulated datasets were used to investigate the effect of the percentage of unknown parentages and numbers of SSR on the accuracy of genealogical coancestries estimated by FT*. Pedigree and molecular data were simulated in the same way as in the previous simulation, except that the selfing rate was not controlled and therefore selfings occurred randomly. We simulated 11 datasets as replicates. Four levels of parentage removal from known lineages (25, 50, 75 and 100 % of the known parentages) and four numbers of SSR (10, 36, 63 and 90) were tested. For each percentage of unknown parentage we conducted eight repetitions, by randomly choosing the parentages to remove, and for each number of SSR we conducted eight repetitions, using the genotype for randomly chosen subsets of SSR. When removing known parentages, random individuals were selected in the pedigree and their two parents were set as missing. FT* was applied to the individuals of the last generation to reconstruct their pedigree from their molecular data and to estimate their genealogical coancestry.

### Testing the effect of LD and HW in the base population with simulated data

The effect of linkage disequilibrium (LD) and departure from Hardy–Weinberg equilibrium (HW) in the base population on the accuracy of genealogical coancestries estimated by FT* was also assessed by simulation.

We first generated a base population departing from HW and with LD ('NON IDEAL' base population) by previously simulating 100 discrete generations with a constant generation size of 20 individuals. In the initial generation, 150 SSR were simulated by randomly drawing without replacement alleles from a pool of 15 equiprobable alleles, independently for each marker.

Sequence repeat markers were evenly distributed along a genome of 10 chromosomes of 150 cM each. Matings were done at random while imposing a minimum percentage of selfings of 20 % per generation, in order to further increase random genetic drift. Each individual had an equal contribution to the next generation. Marker allele segregation was simulated using 'pedantics', with the mutation rate set at zero. The last generation (100) constituted the 'NON IDEAL' base population. From this, we derived an 'IDEAL' population, by randomizing alleles present in the 'NON IDEAL' population within each SSR among the 20 individuals. In this way, the degree of polymorphism and the allelic frequencies remained the same for each marker between the two populations, while they differed by their LD and by the magnitude of the departure from HW equilibrium. LD was measured for each pair of SSR by the squared Pearson correlation ($r^2$) using the R 'gap' package (Zhao 2007). These values were used in a non-parametric paired test of Wilcoxon to check if the mean $r^2$ over all SSR pairs was significantly smaller in the 'IDEAL' base population than in the 'NON IDEAL' base population, i.e., that the simulation successfully created a significantly higher LD in the 'NON IDEAL' base population. Similarly, HW equilibrium was assessed for each SSR with the exact tests of Emigh (1980) for biallelic markers or Guo and Thompson (1992) for multiallelic markers. The $p$ value of each test was kept, with $p$ values lower than 0.05 indicating significant departure from HW. They were used in a non-parametric paired test of Wilcoxon to check if the mean $p$ value over all SSR was significantly smaller in the 'NON IDEAL' base population than in the 'IDEAL' base population, i.e., that the simulation successfully created HW deviation in the 'NON IDEAL' base population. This process was replicated 75 times and only 24 replicates were finally used, as their $p$ value was lower than 0.05 for the two Wilcoxon tests.

Once pairs of 'IDEAL' and 'NON IDEAL' base populations were successfully created, we simulated six additional generations of an expanding pedigree with a random mating regime, including the possibility of selfing. This pedigree was the same for a given pair of 'IDEAL' and 'NON IDEAL' base populations. FT* was applied to the 60 individuals of the last generation to reconstruct their pedigree relationships from the molecular data and to estimate their genealogical coancestry. Consequently, it was possible to compare the accuracy of the method when markers were at equilibrium (both HW and linkage) or when they departed from these 'ideal conditions'.

In order to test the effect of a stronger departure from ideal conditions, this procedure was repeated with a higher imposed minimum percentage of selfing when generating the 'NON IDEAL' base populations (40 %, instead of 20 %). To compensate for the higher random genetic drift associated with the higher selfing rate during the process of creation of the base populations, more SSR were simulated in the initial generation (170, instead of 150). We finally had two simulated datasets, termed 'low disequilibria' and 'high disequilibria', with 24 replicates each.

### Yangambi oil palm population case study with known pedigree

Simulated populations can be considered as somewhat ideal populations compared to real populations where, for example, selection could have been applied. Therefore, FT* was also tested with a real oil palm dataset.

The Yangambi population originated from the Democratic Republic of Congo with the plantation in the 1920s of palms coming from open pollinations of a few founders

considered as unrelated (Corley and Tinker 2003). The subsequent pedigree that started at that point is well known, with a molecular assessment by SSR markers (Cochard 2008) for families that belong to the CIRAD/PalmElit breeding program (http://www.palmelit.com). To validate FT*, we used data from 16 oil palms of these families. Their genealogy covered four generations, back to founder individuals (Fig. 1). This genealogy included selfings from the second generation onwards and one individual was used as male and female in different matings. Moreover, the contribution of individuals to the following generation varied under the effect of artificial selection. The individuals were genotyped with 166 SSR markers.

In order to investigate the effect of the number of SSR on the genealogical coancestries estimated by FT*, we tested seven numbers of SSR (from 6 to 166). For each number of SSR, we conducted eight repetitions, by randomly choosing subsets of SSR.

## Measures of accuracy

In order to measure the accuracy of the pedigree reconstruction when validating the approach with either simulated or real data, the Pearson correlation and the root mean square error (RMSE) were calculated between the true and estimated genealogical coancestries, as described by Fernández and Toro (2006).

## Application to the Deli oil palm population with scarce pedigree data

The Deli population originated from four oil palms planted in 1848 in Indonesia and the first selfings were done in the 1920s (Corley and Tinker 2003). It is now a key oil palm breeding population, as it has a high agronomic value and because its complementation with other populations (in particular La Mé) leads to heterosis in total bunch production (Cochard et al. 2009). Some knowledge on the history of this population is available but detailed genealogical information only exists for the recent past. FT* was applied to 104 oil palms of the last generation of the Deli population used in the CIRAD/PalmElit breeding program. Their pedigree was mostly unknown and the depth of the known part of their pedigree varied among families for up to a maximum of four generations (Fig. S1, available as Supplementary Data). The unknown part of the pedigree of those 104 individuals was reconstructed from their molecular data, back to the four founders. The 104 individuals were genotyped with 160 SSR. In accordance with breeders' knowledge on the history of the population, FT* was

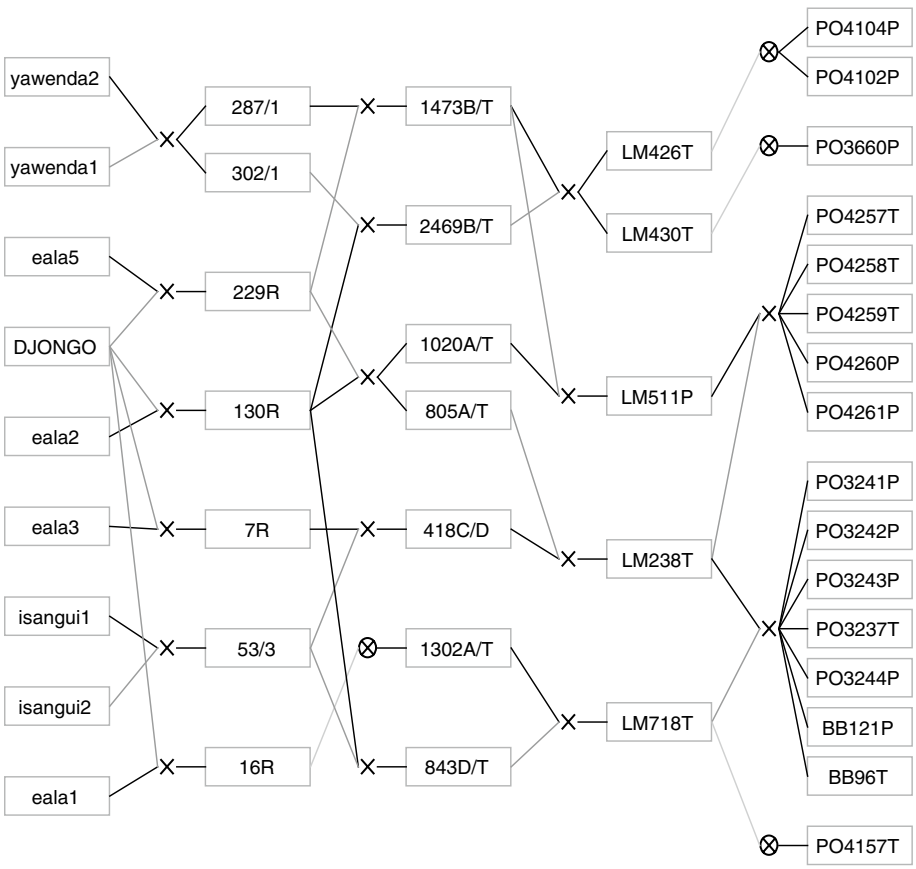


**Fig. 1** Pedigree of the Yangambi population of oil palm (with first generation on the *left* and fourth generation on the *right*). The nine founders were considered unrelated. Pedimap software (Voorrips 2007) was used to produce this figure

run alternatively with seven and nine generations elapsed between founders and studied individuals. In addition, the maximum number of individuals per past generation was set in order not to be limiting for the corresponding number of offspring in the known part of the pedigree. Here, all the genotyped individuals did not exactly belong to the same generation compared to the founders, as in the recent past some families had been submitted to more breeding generations. However, this was not a problem for FT* as it concerned the known part of the pedigree, which was arranged so that the most remote known ancestors of each family were in the same generation compared to founders, i.e., in the pedigree some recent generations were skipped for families submitted to fewer breeding generations (see Fig. S1). We tested the statistical power of the methodology according to the number of SSR using 11 levels of SSR numbers (from 8 to 160). For each number of SSR, 16 repetitions were made, by randomly choosing subsets of SSR. As the true genealogical coancestries were unknown, we only measured the fit between the estimated genealogical coancestries and molecular coancestries with RMSE and Pearson correlation.

When estimating genealogical coancestries, FT* reconstructed a pedigree that was compatible with the observed diversity and the molecular relationship between genotyped individuals. Therefore, the estimated genealogy could be used to calculate historical effective population sizes (*Ne*), which is the size of an idealized Wright–Fisher population that would give rise to the same extent of random genetic drift as the actual population (Caballero 1994). Here we used the reconstructed pedigree to estimate *Ne* with the pedigree-based approach developed by Gutiérrez et al. (2008, 2009) and Cervantes et al. (2011). This approach estimates the realized inbreeding and coancestry *Ne* from the individual increase in inbreeding, which is computed for each individual starting from its most remote ancestors. This *Ne* accounts for differences in the depth of the pedigree between lineages and also for all departures between the real and ideal conditions due, for instance, to selection. A modified version of ENDOG software 4.8 (Gutiérrez and Goyache 2005) which accounts for self-fertilization, was used to calculate realized *Ne* in the Deli breeding population from the reconstructed pedigree. As a control, ENDOG was also applied to the real pedigree of the Yangambi population as well as to its reconstruction (made with all SSR). In order to compare the results with a method independent of the pedigree data, we also used the approach of Hill (1981) and Waples (2006), which calculates inbreeding *Ne* in the parental generation from linkage disequilibrium between unlinked markers. LDNE software version 1.31 (Waples and Do 2008) was used for this task. Calculations were performed with three sets of 16 SSR chosen on different linkage groups, according to the reference map of Billotte et al. (2005).

## Results

### Testing the effect of selfing rate and marker numbers with simulated data

As can be seen in Fig. 2a, the method gave the best results for selfing rates below 0.6 and at least 30 SSR. In these conditions, the root mean square error (RMSE) between the true and estimated genealogical coancestry was small (<0.07). When the selfing rate passed 0.6, the RMSE also increased and finally reached 0.16. The RMSE was the same with 30 or 90 SSR, but was significantly larger with 10 SSR. The evolution of the linear regression line between the estimated and true genealogical coancestries according to the selfing rate helps in analyzing the RMSE profile (Fig. 2b). The decreasing regression slope with increasing selfing rate indicated that the genealogical coancestries were misestimated, with bias increasing with the selfing rate, especially for closely related individuals in the true pedigree. When the selfing rate reached one, the bias became strong as the slope was only 0.55, which coincided with the high observed RMSE.

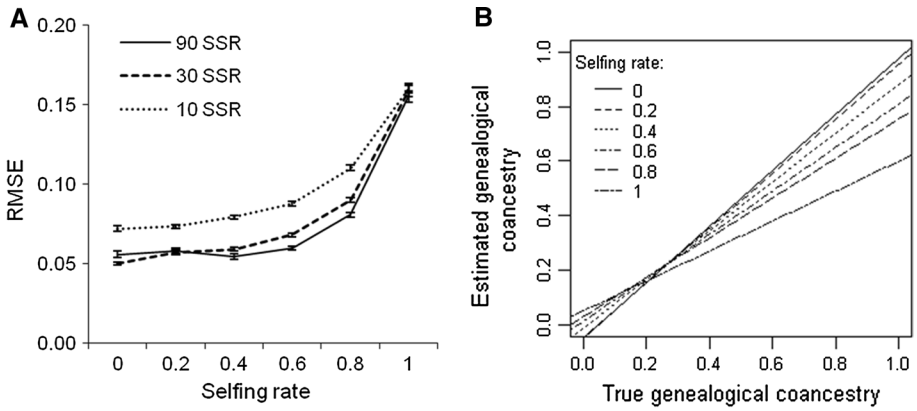


**Fig. 2** Effect of the selfing rate in the true pedigree on **a** the root mean square error (RMSE) between estimated and true genealogical coancestries, according to the number of SSR markers (10, 30 and 90), and **b** on the regression between estimated and true genealogical coancestries using 90 SSR. In **a**, *bars* are SEM ($n = 50$). In **b**, each line is the average regression of estimated coancestry on true genealogical coancestry over 50 replicates

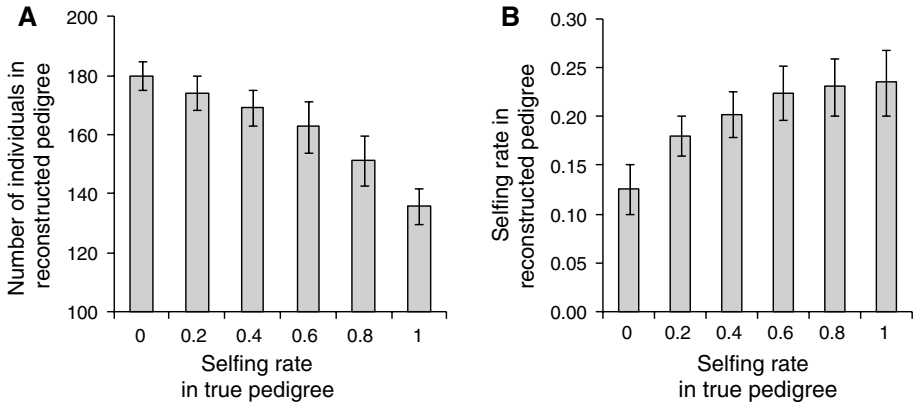


**Fig. 3** Effect of the selfing rate in the true pedigree on **a** the number of individuals (true number of individuals = 120) and **b** the selfing rate in the reconstructed pedigree. 90 SSR were used. *Bars* are SD ($n = 50$)

The above performance was a consequence of the selfing rate and also the interaction of this parameter with the number of virtual individuals in reconstructed pedigrees. The method always overestimated the number of individuals in the reconstructed pedigree, from 13 to 50 % (Fig. 3a) and, therefore, this led to underestimated genealogical coancestries. However, when the true selfing rate was under 0.2, this underestimation of coancestries was partly compensated by an overestimation of the selfing rate (Fig. 3b), leading to a small RMSE. When the true selfing rate increased above 0.2, FT* underestimated the selfing rate, as it was unable to generate pedigrees with a selfing rate higher than 0.30, leading to a greater coancestry underestimation.

Surprisingly, the Pearson correlation between the real and estimated genealogical coancestries gave apparently contradictory results, as it increased in parallel with the selfing rate (Fig. S2, available as Supplementary Data). However, the Pearson correlation was actually not relevant to evaluate the effect of an increase in the self-fertilization level. Indeed, higher selfing rates led to more homogeneous and differentiated families, so the estimates had high correlations even when the values were biased (high RMSE). This is illustrated in Fig. S3, available as Supplementary Data.

## Testing the effect of the percentage of unknown parentages and marker numbers with simulated data

According to the number of markers used and the percentage of parentages removed, the Pearson correlation between the true and estimated coancestries ranged from 0.8 to 0.969 and the RMSE from 0.033 to 0.074 (Fig. 4a, b). The Pearson correlation increased and the RMSE decreased with the number of markers, especially between 10 and 36 SSR and little change in both parameters was observed with a further increase in the number of markers. As expected, the Pearson correlation decreased and the RMSE increased when a larger part of the pedigree

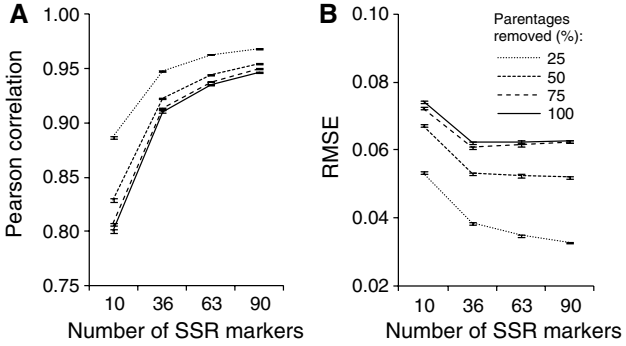


**Fig. 4** Effect of the percentage of unknown parentages and number of SSR markers on the **a** Pearson correlation and **b** root mean square error (RMSE) between estimated and true genealogical coancestries. *Bars* are SEM ($n = 704$)

was unknown. When the pedigree was assumed to be completely unknown, the RMSE plateaued at 36 SSR and the Pearson correlation at 90 SSR. Using sets of 30–90 SSR led to a Pearson correlation of over 0.9 and an RMSE under 0.07.

## Testing the effect of LD and HW in the base population with simulated data

The mean $r^2$ measure of LD in the 'NON IDEAL' base populations was 48.4 % higher than in the 'IDEAL' populations in the 'low disequilibria' datasets and 60 % higher in the 'high disequilibria' datasets ($p < 0.001$ for all replicates). The mean $p$ value of the HW test in the 'IDEAL' base populations was 8.8 % higher than in the 'NON IDEAL' populations in the 'low disequilibria' datasets and 41.5 % higher in the 'high disequilibria' datasets ($p < 0.05$ for all replicates) (Table 2). As random genetic drift induced allele fixation at some SSR during the process of creation of the base populations, the final number of SSR that were polymorphic in the base population and used for pedigree reconstruction fell to an average of 92 ± 6 (SD)

and 103 ± 8 in the 'low' and 'high disequilibria' datasets, respectively. In the last generation (genotyped individuals whose pedigree was reconstructed), polymorphism for these SSR was low, with an average of 2.2 ± 0.4 (SD) and 2.1 ± 0.4 in the 'low' and 'high disequilibria' datasets, respectively. Finally, the simulation results showed that the status of the base population regarding the HW and linkage equilibrium had no effect on the genealogical coancestry estimation, as both the RMSE and Pearson correlation were similar with the two base populations, regardless of the strength of the departure from the ideal conditions (Table 2).

### Yangambi oil palm case study with known pedigree

The Pearson correlation increased and the RMSE decreased with the number of markers, especially between 6 and 38 SSR, and little change in both parameters was observed with a further increase in the number of markers (Fig. 5a, b). Using 38 SSR led to a Pearson correlation of above 0.9 and an RMSE under 0.08. This result was in agreement with the simulation results.

Selfing occurred in this population at a rate of 0.18, while the rate estimated from the reconstructed pedigree was 0.27 when using all markers. This discrepancy in the selfing rate between the true and reconstructed pedigree was consistent with the simulation results obtained with a true selfing rate close to 0.20. The number of individuals in the pedigree was overestimated by 34.1 %, which was also consistent with the simulation results.

When using FT* with a real dataset, the only summary statistic is the Pearson correlation between the molecular and estimated genealogical coancestries. To check if this statistic could be used as a measure of error of the true pedigree, we compared its evolution with the evolution of RMSE and Pearson correlation between the true and estimated genealogical coancestries according to the number of SSR. The Pearson correlation between the molecular and estimated genealogical coancestries was high even with 6 SSR (0.95 ± 0.01 SD) and reached a plateau at 70 SSR (not shown). Therefore, when applying FT* the effect of the number of markers must be investigated in order to know if enough markers were used to achieve the best possible result that the method can yield for the dataset. This point is crucial, as a small increase in the Pearson correlation between the molecular and estimated genealogical coancestries can be associated with a strong increase in the quality of the pedigree reconstruction. Thus, in the Yangambi

**Table 2** Effect of the base population on genealogical coancestry estimation

| Base population | $r^2$ (LD) | $p$ value (HW) | RMSE | Pearson |
|---|---|---|---|---|
| IDEAL | 0.031 ± 0.002 | 0.626 ± 0.043 | 0.112 ± 0.019 | 0.739 ± 0.038 |
| NON IDEAL (low disequilibria) | 0.046 ± 0.003 | 0.576 ± 0.048 | 0.109 ± 0.018 | 0.742 ± 0.035 |
| IDEAL | 0.035 ± 0.002 | 0.638 ± 0.024 | 0.099 ± 0.015 | 0.777 ± 0.022 |
| NON IDEAL (high disequilibria) | 0.056 ± 0.004 | 0.451 ± 0.061 | 0.104 ± 0.017 | 0.754 ± 0.029 |

Values are mean ± SD ($n = 24$). $r^2$ is the measure of linkage disequilibrium (LD) obtained by the squared Pearson correlation between markers. $p$ Value (HW) comes from exact tests for Hardy–Weinberg departure. The root mean square error (RMSE) and the Pearson correlation were calculated between estimated and true genealogical coancestries

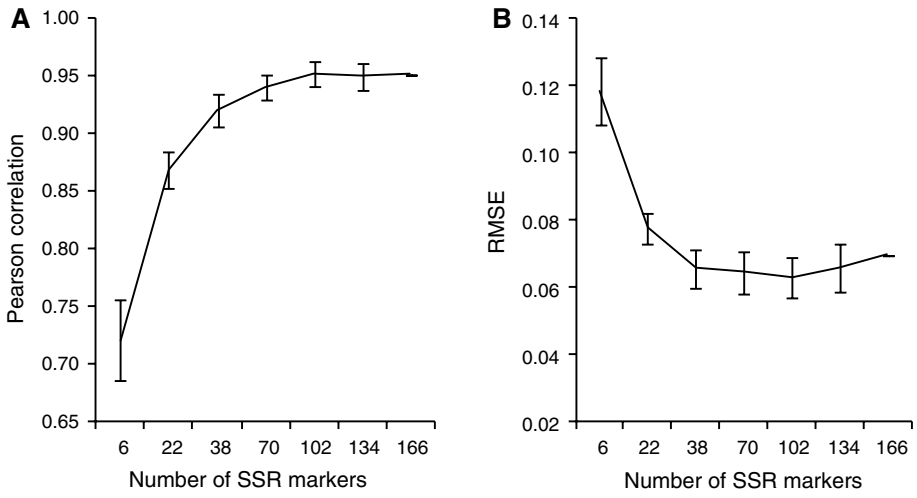


**Fig. 5** Effect of the number of SSR markers on the **a** Pearson correlation and **b** root mean square error (RMSE) between estimated and true genealogical coancestries in the Yangambi oil palm breeding population. *Bars* are SEM ($n = 8$)

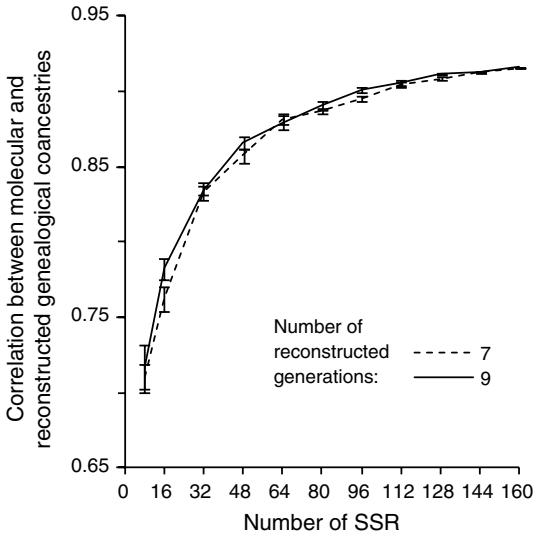


**Fig. 6** Effect of the number of SSR markers and the number of generations to reconstruct on the Pearson correlation between molecular and estimated genealogical coancestries in the Deli oil palm breeding population. *Bars* are SEM ($n = 16$)

dataset, while the Pearson correlation between the molecular and estimated genealogical coancestries increased from 0.95 to 0.98, the Pearson correlation between the true and estimated genealogical coancestries increased from 0.72 to 0.95 and the RMSE decreased from 0.12 to 0.07.

### Application to the Deli oil palm population with scarce pedigree data

According to the number of markers used and the depth of the reconstructed pedigree, the Pearson correlation between the molecular and estimated genealogical coancestries ranged from 0.710 to 0.916 (Fig. 6). The Pearson correlation increased with the number of SSR according to a diminishing return trend. For all numbers of SSR, the Pearson correlation was similar, with seven and nine reconstructed generations, with slightly higher levels for the latter case but with no significant differences. Fernández and Toro (2006) already found that the correlation between the true and estimated genealogical coancestries increased with the number of virtual generations, even when the true genealogy had less generations than the estimated one. A high Pearson correlation between the estimated genealogical coancestries and molecular coancestries (>0.9) was achieved when using more than 100 markers.

For seven and nine generations, the realized inbreeding *Ne* was 4.3 ± 1.3 (SD) and 5.3 ± 1.3 and the ratios of the coancestry *Ne* to the inbreeding *Ne* were 2.0 and 1.8, respectively. LDNE gave an inbreeding *Ne* of 5.0 ± 1.1. In the Yangambi population, the realized inbreeding *Ne* from the reconstructed pedigree (10.9 ± 4.3) and the true pedigree (7.1 ± 2.4) were not significantly different, indicating that using the reconstructed pedigree to estimate *Ne* was relevant.

## Discussion

When the pedigree is unknown or scarcely known, it could be worthwhile to recover or estimate the genealogical relationship between available individuals on the basis of molecular information. Fernández and Toro (2006) method (FT) exhibits good properties compared to other methods reported in the literature. However, FT was originally designed to deal with data from species with separate sexes and, therefore, it was not applicable to many plant species. In the present study, we have adapted the FT method to deal with hermaphroditism and monoecy, with the possibility of selfing. In addition, some improvements were made to the method to take previous knowledge on the population demographic history into account. The new version of the method (FT*) showed good performance on simulated as well as real data of mixed-mating species using a realistic number of markers. The influence of different parameters on the accuracy of the method was also determined.

### Accuracy and current limitations of the modified FT method

The FT* method overestimated the number of individuals in the genealogy, while the selfing rate could be either over- or underestimated, depending on the true selfing rate. The best results were obtained for a selfing rate of below 0.6, which shows that this method is suitable for many plant populations. For instance, Jarne and Auld (2006) reported that around 60 % of hermaphroditic plant species had a selfing rate below 0.4. However, when the true selfing rate is likely to be very high or very low, it would be interesting to use this information as a constraint in the pedigree reconstruction. Consequently, further improvements of the method could include a user-specified percentage of selfing for any of the virtual ancestor generations.

With our simulated data, the new method gave similar results to those presented in Fernández and Toro (2006) for unbalanced simulated datasets of small populations. When using enough markers, they also achieved a small RMSE. Moreover, our real Yangambi oil palm data and the real pig data used by Fernández and Toro (2006) yielded similar results. As we could expect, both datasets showed that the number of markers was a limiting factor, since the reliability of the coancestry estimates was markedly reduced when it was too small.

Marker polymorphism appeared to be another key parameter interacting with the number of markers required

to get the best results. In the pig dataset used by Fernández and Toro (2006), one marker per chromosome with 4.2 alleles per marker gave results very close to the best values obtained (using all markers), which was also the case in our first and second sets of simulations using only 10 SSR. In our third group of simulations, although an average number of 92 SSR were used, the coancestry estimation did not achieve the same quality as in the first and second simulations, as indicated by the higher RMSE and lower Pearson correlation. This could be related to the marker polymorphism, which was lower in the third simulation (an average of 2.2 alleles per SSR). Finally, when the number of markers was sufficiently high (from 30 to 100, according to their polymorphism), a high Pearson correlation and low RMSE between true and estimated coancestries were obtained.

The maximum number of individuals in past generations also played a very important role in the quality of the coancestry estimation. An artificially enlarged number of ancestors leads to a larger feasible space of solutions thus making it harder for the underlying optimization algorithm to find the fittest solution. Here, the maximum number of individuals per past generation was at least twofold higher than the true value, which led FT* to substantially overestimate the number of individuals. Clearly, if the maximum number of past individuals could be accurately defined (roughly, less than one-third higher than their true number), the overestimation of the number of individuals and, consequently, the selfing rate bias would have been lower. All of these considerations highlight the importance of including correct information on the past demographic structure of the population to get accurate estimates, as close as possible to the true pedigree.

As FT* makes no assumptions about Hardy–Weinberg and linkage equilibria in the base population, its results were expected to be unaffected by possible departure from these ideal conditions. We confirmed this point here. This is a very important feature of FT, as many methods used to infer relationships from molecular markers, either by explicit pedigree reconstruction or using pairwise estimators, make these assumptions (Fernández and Toro 2006). Moreover, as the number of markers increases rapidly (through SNP panels or Next Generation Sequencing), it will be common to have markers in linkage disequilibrium.

The FT* method is only suitable for diploid species. It had not been a concern in the original version as polyploid animal species are rare. It is however much more common in plant species, as many important crops are polyploid. Clearly, the current version could be extended to polyploid species. The method to calculate molecular coancestries should be modified and changes should be made in the rules to check for molecular incompatibilities within full-sib families, but this would be rather straightforward. From an operational standpoint, the program should account for individuals carrying more than two alleles per loci and adequately simulate the way of transmission of genetic information from one generation to the next.

As stated before, FT* only considers discrete generations of virtual ancestors. Thus, it assumes that the parents of an individual always belong to the previous generation and that mating between two individuals of different generations is not possible. This point can be limiting, for instance for natural populations of perennial species or for breeding populations where clonally propagated individuals have been used repeatedly throughout the pedigree. Notwithstanding, the objective of the method is to estimate a genealogical coancestry matrix between the available individuals at a particular time (e.g., selection candidates in a breeding program), not necessarily to reconstruct the exact real pedigree leading to them. Consequently, the aim is to get a pedigree which is compatible with the observed structure of molecular relationships and that implies the same level of genetic drift. Another limitation of the method is that it assumes that all available individuals with a genotype belong to the last generation, while molecular data can also be available for individuals of past generations even if they are no longer present in the population. To account for both situations, the objective would be somewhat different, i.e., to reconstruct the true pedigree, at least for all genotyped individuals. Modifying the method to allow for the use of molecular data over several generations would require checking the molecular compatibility between relationships apart from full-sib families and would also require using other information such as the date of birth of each individual or the age when they were reproductively mature. This task would be computationally more complex than in the present situation.

Few methods explicitly reconstruct a multigeneration pedigree from molecular data. Almudevar (2003) and Riester et al. (2009) used a simulated annealing algorithm to find the maximum likelihood pedigree. Cowell (2009) developed an exhaustive search algorithm adapted from a Bayesian network learning algorithm. Almudevar (2007) used a computationally intensive fully Bayesian approach to infer a pedigree graph from molecular data. However, all these methods applied to complete samples of individuals, i.e., that all the individuals appearing in the final reconstructed pedigree had molecular data, and possible unsampled parents were assumed unrelated to the others. To our knowledge, only two approaches [FT* and the method of Gasbarra et al. (2007a, b)] apply to genotyped individuals belonging to a single generation. Like FT*, the method of Gasbarra et al. (2007a, b) reconstructs a pedigree from molecular data of contemporaneous individuals and information about the population history (number of generations from founders and approximate size of the population at each generation), assuming nonoverlapping

generations. It differs from FT* as it is based on a Bayesian approach using a Markov chain Monte Carlo algorithm. An interesting feature of this method is that it models both the pedigree and gene flows from the founders down to the genotyped individuals. This ensures molecular compatibility at the pedigree level, not only within full-sib families. However, it requires more data than FT*. Mating parameters controlling the distribution of offspring among males and the degree of monogamy (or estimated from the data), as well as allele frequencies in the base population must be known. A further study is needed to compare the results of those two approaches for the same dataset.

### Future uses of FT*

In our study, we focused on breeding populations as we considered that the method would give the best results in these situations. Indeed, their pedigree is generally only partly missing and the known part of the pedigree reduces the possibilities, thus helping the FT* method to approximate the correct relationships. Furthermore, a lot of other information is often available and would further reduce the number of feasible solutions. For example, the structures of some breeding populations contain just full-sib families, or the maximum number of male parents can be established, for instance in polycross designs. However, FT* could also be applied to natural populations. This was already the case with the original FT method, which was applied to several natural animal populations. For example, Zub et al. (2012) estimated heritability in a wild weasel population by applying an animal model using the pedigree reconstructed by FT. Therefore, FT* could be used to study the genetic structure of natural plant populations, evaluate the genetic diversity they harbor and/or design effective management strategies. The FT* method could also be of help in the management of conserved populations, as Fernández et al. (2005) showed that combining molecular markers and pedigree information was useful for increasing the effective population size. However, investigations should be conducted to determine whether combining the molecular coancestry and the genealogical coancestry estimated from markers could give better results than using the molecular coancestry alone, as they could provide redundant information.

### Application to the Deli oil palm population with scarce pedigree data

Our use of the reconstructed pedigree of the Deli oil palm breeding population was similar to the approach implemented by Cervantes et al. (2011) in animals, where the reconstructed pedigree of ruminant populations was used to estimate their realized effective size. Here, we found that the reconstructed pedigree was appropriate to estimate *Ne*. However, we constrained FT* with three items of genealogical information: number of founders, known pedigree in recent generations and approximate number of generations between founders and genotyped individuals. Therefore, in this case the reconstructed pedigree should be rather close to the true pedigree, and pedigree-based statistics should be reliable. When FT* is used with no prior information about the history of the population, users should be cautious with statistics calculated with the reconstructed pedigree.

The estimation of *Ne* through the realized inbreeding *Ne* and LD methods refer to different periods of time. The realized inbreeding *Ne* is an average over the time period covered by the pedigree, while LDNE gives the *Ne* of the parental generation. The bottleneck event at the founding of the Deli population created LD, which declined thereafter as the population expanded. Therefore, we could expect the realized *Ne* to be lower than the LDNE value. However, Waples (2005) showed that *Ne* based on LD can be underestimated for several generations under the effects of recent bottlenecks and population size increases. Therefore, the LDNE value could have actually been affected by the recent history of the population, and also reflect the *Ne* before the parental generation. In any case, the results obtained with these two methods in our data appeared to be consistent with each other.

This is the first report of *Ne* in oil palm. Meuwissen (2009) estimated that the critical *Ne* for a population was around 50–100, on order to avoid long-term inbreeding problems. Our results highlighted that efforts should be made to limit inbreeding in the oil palm breeding populations studied and to diversify their genetic base. For example, this could be achieved for the Deli population by crossing it with some African populations, as suggested by Cochard et al. (2009). As the coancestry *Ne* and the inbreeding *Ne* are measures of the same drift process, they should reach an identical asymptotic value in an idealized population without permanent sublining. Therefore, their difference gives an estimate of preferential matings. In the Deli population, the ratio of the coancestry *Ne* to the inbreeding *Ne* indicated a high degree of subdivision due to nonrandom mating. This could likely be explained by the selection applied to this population, which first underwent mass selection and reciprocal recurrent selection afterwards. This was also reflected in the high variability in the average relatedness of the four founders, which is a measure of their genetic contribution to the population (Gutiérrez and Goyache 2005). For instance with seven reconstructed generations, the average relatedness given by ENDOG ranged from 8.3 to 44.2 %.

## Conclusion

The FT* method gave reliable coancestry estimates for mixed-mating species, especially when the selfing rate was lower than 0.6, using a realistic number of markers. We confirmed that the existence of linkage disequilibrium and departure from the Hardy–Weinberg equilibrium in the base population did not affect the method. In a case study, this approach gave valuable information about the Deli oil palm population. This highlighted the potential benefits that plant breeders could obtain by looking for new tools in the animal breeding sector and adapting them to their own circumstances. The method was implemented in the software MOLCOANC 3.0, where the user can choose between separate sexes and mixed-mating. The program is available from the web page http://dl.dropbox.com/u/5714008/Fernandez.htm.

**Acknowledgments** We would like to thank Virginie Pomiès (CIRAD, Montpellier) for her technical assistance in genotyping, Dr Juan Pablo Gutiérrez (Universidad Complutense de Madrid) for his help with ENDOG software and the anonymous reviewers for their helpful comments. This research was partly funded by a grant from PalmElit SAS.

**Conflict of interest** The authors declare no conflict of interest.

## References

Adams W, Neale D, Loopstra C (1988) Verifying controlled crosses in conifer tree-improvement programs. Silvae genetica 37:147–152

Almudevar A (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor Popul Biol 63:63–75

Almudevar A (2007) A graphical approach to relatedness inference. Theor Popul Biol 71:213–229

Atkin FC, Dieters MJ, Stringer JK (2009) Impact of depth of pedigree and inclusion of historical data on the estimation of additive variance and breeding values in a sugarcane breeding program. Theor Appl Genet 119:555–565

Billotte N, Marseillac N, Risterucci AM et al (2005) Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 110:754–765

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evolut Pers Ed 18:503–511

Butler K, Field C, Herbinger CM, Smith BR (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. Mol Ecol 13:1589–1600

Caballero A (1994) Developments in the prediction of effective population size. Heredity 73:657–679

Cervantes I, Goyache F, Molina A, Valera M, Gutiérrez JP (2011) Estimation of effective population size from the rate of coancestry in pedigreed populations. J Anim Breed Genet 128:56–63

Cochard B (2008) Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (*Elaeis guineensis* Jacq.). Montpellier SupAgro, Montpellier, p 175

Cochard B, Adon B, Rekima S et al (2009) Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. Tree Genetics Genomes 5:493–504

Corley RHV (2005) Illegitimacy in oil palm breeding—a review. J Oil Palm Res 17:64–69

Corley RHV, Tinker PB (2003) Selection and breeding. In: Blackwell Science Ltd (ed) The oil palm, 4th edn. Blackwell Publishing, Oxford, pp 133–199

Cowell RG (2009) Efficient maximum likelihood pedigree reconstruction. Theor Popul Biol 76:285–291

Doerksen TK, Herbinger CM (2010) Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. Tree Genetics Genomes 6:591–600

Eding H, Meuwissen THE (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159

Emigh T (1980) Comparison of tests for Hardy–Weinberg equilibrium. Biometrics 36:627–642

Emik LO, Terrill CE (1949) Systematic procedures for calculating inbreeding coefficients. J Hered 40:51–55

Ericsson T (1999) The effect of pedigree error by misidentification of individual trees on genetic evaluation of a full-sib experiment. Silvae genetica 48:239–242

Fernández J, Toro MA (2006) A new method to estimate relatedness from molecular markers. Mol Ecol 15:1657–1667

Fernández J, Villanueva B, Pong-Wong R, Toro MA (2005) Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170:1313–1321

Garbe JR, Da Y (2008) Pedigraph user manual Version 2.4. Department of Animal Science, University of Minnesota

Gasbarra D, Pirinen M, Sillanpaa M, Arjas E (2007a) Estimating genealogies from linked marker data: a Bayesian approach. BMC Bioinforma 8:411

Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007b) Estimating genealogies from unlinked marker data: a Bayesian approach. Theor Popul Biol 72:305–322

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372

Gutiérrez JP, Goyache F (2005) A note on ENDOG: a computer program for analysing pedigree information. J Anim Breed Genet 122:172–176

Gutiérrez JP, Cervantes I, Molina A, Valera M, Goyache F (2008) Individual increase in inbreeding allows estimating effective sizes from pedigrees. Genet Sel Evol 40:359–378

Gutiérrez JP, Cervantes I, Goyache F (2009) Improving the estimation of realized effective population sizes in farm animals. J Anim Breed Genet 126:327–332

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. Genet Res 38:209–216

Jarne P, Auld JR (2006) Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution 60:1816–1824

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Kumar S, Gerber S, Richardson TE, Gea L (2007) Testing for unequal paternal contributions using nuclear and chloroplast SSR markers in polycross families of radiata pine. Tree Genetics Genomes 3:207–214

McIntyre CL, Jackson PA (2001) Low level of selfing found in a sample of crosses in Australian sugarcane breeding programs. Euphytica 117:245–249

Meuwissen T (2009) Genetic management of small populations: a review. Acta Agriculturae Scandinavica Section A Animal Sci 59:71–79

Morrissey MB, Wilson AJ (2010) pedantics: an r package for pedigree-based genetic simulation and pedigree manipulation, characterization and viewing. Mol Ecol Resour 10:711–719

Pemberton JM (2008) Wild pedigrees: the way forward. Proc Royal Soc B Biol Sci 275:613–621

Riester M, Stadler PF, Klemm K (2009) FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics 25:2134–2139

Voorrips RE (2007) Pedimap: software for visualization of genetic and phenotypic data in pedigrees. Plant Research International, Wageningen

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? Mol Ecol 14:3335–3352

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv Genet 7:167–184

Waples RS, Do CHI (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour 8:753–756

Zhao J (2007) gap: genetic analysis package. J Stat Softw 23:1–18

Zub K, Piertney S, Szafranska PA, Konarzewski M (2012) Environmental and genetic influences on body mass and resting metabolic rates (RMR) in a natural population of weasel *Mustela nivalis*. Mol Ecol 21:1283–1293